LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Gerhard Tutz & Andreas Groll

Binary and Ordinal Random Effects Models Including Variable Selection

Technical Report Number 097, 2010
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Binary and Ordinal Random Effects Models Including Variable Selection

Gerhard Tutz * Andreas Groll †

December 10, 2010

Abstract

A likelihood-based boosting approach for the fitting of binary and ordinal mixed models is presented. In contrast to common procedures it can be used in high-dimensional settings where a large number of potentially influential explanatory variables is available. Constructed as a componentwise boosting method it is able to perform variable selection with the complexity of the resulting estimator being determined by information criteria. The method is investigated in simulation studies both for cumulative and sequential models and is illustrated by using real data sets.

Keywords Binary mixed model, Ordinal mixed model, Cumulative model, Sequential model, Boosting, Variable selection, Penalized Quasi-Likelihood, Laplace approximation

*Department of Statistics, University of Munich, Akademiestrasse 1, D-80799, Munich, Germany, *email*: tutz@stat.uni-muenchen.de

†Department of Statistics, University of Munich, Akademiestrasse 1, D-80799, Munich, Germany, *email*: andreas.groll@stat.uni-muenchen.de

1 Introduction

Random effects models are a common tool for the modeling of heterogeneity and dependence of responses in repeated measurements. We will focus on the ordinal response case which includes binary responses. For ordinal response variables various models and estimation methods have been proposed. Jansen (1990) and Tutz and Hennevogl (1996) considered cumulative type random effects models, adjacent categories type models were considered by Hartzel et al. (2001). However, the proposed methods apply only if few explanatory variables are included. In particular for cumulative type models existence of parameter estimates is a problem if a large number of explanatory variables is available and no selection procedure is used. Forward selection procedures, which can be constructed for generalized linear models and extensions to random effects models, have the disadvantage of being rather unstable.

One way to reduce the predictor space by variable selection is to use penalization techniques, which are widely available for generalized linear models (GLMs) but not for random effects models. Examples are the lasso for GLMs (Park and Hastie, 2006) or SCAD (Fan and Li, 2001). An alternative method uses boosting methods which have been developed in the machine learning community. The stepwise fitting procedure, which improves only selected coefficients within one step, implicitly selects predictors. But most procedures work for the linear model or GLMs. In the present paper boosting algorithms are proposed that select variables in the mixed model framework. The estimation methods necessarily differ from common boosting procedures for GLMs because the mixture component has to be adapted within the stepwise fitting procedure. We will consider two types of ordered models and different fitting procedures.

In Section 2 a short review of ordinal random effects models is given. In Section 3 the fitting procedure is outlined and some simulation results that include comparison to established estimation methods are reported. Applications are found in Section 4.

2 Ordinal Random Effects Models

A frequently encountered type of data is where the response variables are measured on an ordinal scale. The following models for repeatedly assessed ordinal scores are based on the threshold concept, which means that the observed category is assumed to be determined by the value of a latent unobservable continuous response.

2.1 Models for Ordinal Response Variables

Several models for ordinal response variables are in common use. The most widely used model is the cumulative model, which was propagated by McCullagh (1980). With the response Y_i taking values from $\{1, \dots, k\}$ the cumulative model has the form

$$P(Y_i \leq r|\mathbf{x}_i) = F(\gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma}), \quad r = 1, \dots, k \tag{1}$$

where $-\infty = \gamma_{00} \leq \gamma_{01} \leq \cdots \leq \gamma_{0k} = \infty$ are category-specific intercepts. The model may be derived from an underlying latent regression model which has a noise term with distribution function F. Then the categorical response Y is a coarser version determined by the ordered thresholds $\gamma_{00}, \dots, \gamma_{0k}$ (see for example McCullagh, 1980). For the logistic distribution function the cumulative model is known as the proportional odds model. Alternative models are the sequential model

$$P(Y_i = r|Y_i \geq r, \mathbf{x}_i) = F(\gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma}), \quad r = 1, \dots, q := k - 1$$

and the adjacent category model

$$P(Y_i = r|Y_i \in \{r, r+1\}) = F(\gamma_{0r} + \mathbf{x}_i^T\boldsymbol{\gamma}_r), \quad r = 1, \dots, q,$$

where F again is a strictly monotone distribution function and $q = k - 1$. An advantage of the sequential and adjacent type model is that no order restriction on intercepts is needed.

All of these models can be given in matrix form as multivariate generalized linear models (GLMs) for categorical responses. For observation i one obtains

$$g(\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\beta} \quad \text{or} \quad \boldsymbol{\pi}_i = h(\mathbf{X}_i\boldsymbol{\beta}),$$

where $\boldsymbol{\pi}_i^T = (\pi_{i1}, \ldots, \pi_{iq})$, $\pi_{ir} = P(Y_i = r|\mathbf{x}_i)$, is the vector of response probabilities, g is the (multivariate) link function, $h = g^{-1}$ is the inverse link function and $\mathbf{X}_i$ is an appropriate design matrix. The components of $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta}$ have the form

$$\eta_{ir} = \gamma_{0r} + \mathbf{x}_{ir}^T\boldsymbol{\gamma} \qquad \text{or} \qquad \eta_{ir} = \gamma_{0r} + \mathbf{x}_{ir}^T\boldsymbol{\gamma}_r, \quad r = 1, \ldots, q.$$

(for details see Fahrmeir and Tutz, 2001).

2.2 Incorporation of Random Effects

For clustered data let the ordinal response $Y_{it} \in \{1, \ldots, k\}$ denote measurement t in cluster $i, i = 1, \ldots, n$, $t = 1, \ldots, T_i$. In random effects models one assumes that the corresponding model for observation Y_{it} has the form

$$g(\boldsymbol{\pi}_{it}) = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{Z}_{it}\mathbf{b}_i,$$

where $\boldsymbol{\pi}_{it}^T = (\pi_{it1}, \ldots, \pi_{itq})$ denotes the vector of response probabilities with $\pi_{itr} = P(Y_{it} = r|\mathbf{X}_{it}, \mathbf{Z}_{it}, \mathbf{b}_i)$, $\boldsymbol{\beta}$ is a fixed coefficient vector and $\mathbf{b}_i$ is a s-dimensional cluster-specific random effect, for which a distribution, for example $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{Q})$ is assumed. The corresponding linear predictor $\boldsymbol{\eta}_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{Z}_{it}\mathbf{b}_i$ has components

$$\eta_{itr} = \gamma_{0r} + \mathbf{x}_{it}^T\boldsymbol{\gamma} + \mathbf{z}_{it}^T\mathbf{b}_i.$$

The simplest random effects model is a model that includes random intercepts only. It has linear predictor $\eta_{itr} = \gamma_{0r} + \mathbf{x}_{it}^T\boldsymbol{\gamma} + b_i$, where $b_i \sim N(0, \sigma^2)$. Thus each cluster is assumed to have its own response level. More general one might assume that all effects are subject-specific by assuming $\eta_{itr} = \gamma_{0r} + \mathbf{x}_{it}^T\boldsymbol{\gamma} + \mathbf{x}_{it}^T\mathbf{b}_i$.

The corresponding vectors $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are easily derived. The $q \times (q + p)$ -dimensional matrix $\mathbf{X}_{it}$ typically has the form

$$\mathbf{X}_{it} = \begin{pmatrix} 1 & & & \mathbf{x}_{it}^T \\ & 1 & & \\ & & \ddots & \vdots \\ & & & 1 \quad \mathbf{x}_{it}^T \end{pmatrix} = [\mathbf{I}_q, \mathbb{1}_q \otimes \mathbf{x}_{it}^T],$$

with parameter vector $\boldsymbol{\beta}^T = (\gamma_{01}, \ldots, \gamma_{0q}, \boldsymbol{\gamma}^T)$. The form of $\mathbf{Z}_{it}$ depends on the structure of the cluster-specific effects.

For the t-th observation of cluster i we use the notation Y_{it}, which takes values from $\{1, \ldots, k\}$, or $\mathbf{y}_{it}^T = (y_{it1}, \ldots, y_{itq})$, where $y_{itr} = 1$ if $Y_{it} = r$ and $y_{itr} = 0$ otherwise. Then the vector of probabilities is the mean $\boldsymbol{\pi}_{it} = E(\mathbf{y}_{it})$. The observations can be summarized as $(\mathbf{y}_{it}, \mathbf{x}_{it}, \mathbf{z}_{it})$, $i = 1, \ldots, n$, $t = 1, \ldots, T_i$. Observations of one cluster can be combined to yield

$$g(\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i,$$

where $\mathbf{Z}_i^T = [\mathbf{Z}_{i1}^T, \ldots, \mathbf{Z}_{iT_i}^T]$, $\mathbf{X}_i^T = [\mathbf{X}_{i1}^T, \ldots, \mathbf{X}_{iT_i}^T]$ and $\boldsymbol{\pi}_i^T = (\boldsymbol{\pi}_{i1}^T, \ldots, \boldsymbol{\pi}_{iT_i}^T)$. For all observations one obtains

$$g(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b},$$

where $\boldsymbol{\pi}^T = (\boldsymbol{\pi}_1^T, \ldots, \boldsymbol{\pi}_n^T)$, $\mathbf{X}^T = [\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T]$ and $\mathbf{Z} = diag(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is a block-diagonal matrix. For the random effects vector $\mathbf{b}^T = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)$ one assumes a normal distribution with block-diagonal matrix $\mathbf{Q}_\mathbf{b} = diag(\mathbf{Q}, \ldots, \mathbf{Q})$.

Since the multinomial distribution is from the exponential family the models can be embedded into the framework of multivariate generalized linear mixed models (GLMMs). The conditional density of $\mathbf{y}_{it}$, given explanatory variables and the random effect $\mathbf{b}_i$, is of exponential family type

$$f(\mathbf{y}_{it}|\mathbf{X}_{it}, \mathbf{b}_i) = \exp\left\{\frac{(\mathbf{y}_{it}^T\boldsymbol{\theta}_{it} - \kappa(\boldsymbol{\theta}_{it}))}{\phi} + c(\mathbf{y}_{it}, \phi)\right\}, \tag{2}$$

where $\boldsymbol{\theta}_{it} = \theta(\boldsymbol{\pi}_{it})$ denotes the natural parameter, $\kappa(\boldsymbol{\theta}_{it})$ is a specific function corresponding to the type of exponential family, $c(.)$ the log normalization constant and ϕ the dispersion parameter, which in the present case of single multinomial responses is fixed as $\phi = 1$ (compare Fahrmeir and Tutz, 2001).

2.3 Estimation of Ordinal Random Effects Models

Estimates of random effects models can be obtained in various ways. When the dimension of the random effect is low Gauss-Hermite and Monte Carlo methods can be used (for example Tutz and Hennevogl, 1996). An alternative approach that can be extended to multivariate generalized linear mixed models is penalized quasi-likelihood (PQL), which has been suggested by Breslow and Clayton (1993), Lin and Breslow (1996) and Breslow and Lin (1995). Let the covariance matrix $\mathbf{Q}(\boldsymbol{\varrho})$ of the random effects $\mathbf{b}_i$ depend on an unknown parameter vector $\boldsymbol{\varrho}$ which specifies the correlation. In penalization-based concepts the joint likelihood-function is specified by the parameter vector $\boldsymbol{\varrho}$ and parameter vector $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \mathbf{b}^T)$. With $\mathbf{y}_i^T = (\mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{iT_i}^T)$ we obtain the corresponding log-likelihood

$$l(\boldsymbol{\delta}, \boldsymbol{\varrho}) = \sum_{i=1}^{n} \log\left(\int f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\varrho}) p(\mathbf{b}_i, \boldsymbol{\varrho}) d\mathbf{b}_i\right), \tag{3}$$

where $p(\mathbf{b}_i, \boldsymbol{\varrho})$ denotes the density of the random effects. Approximation of (3) along the lines of Breslow and Clayton (1993) yields the penalized likelihood

$$l^{\text{pen}}(\boldsymbol{\delta}, \boldsymbol{\varrho}) = \sum_{i=1}^{n} \log(f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\varrho})) - \frac{1}{2}\mathbf{b}^T\mathbf{Q}(\boldsymbol{\varrho})^{-1}\mathbf{b}, \tag{4}$$

where the penalty term $\mathbf{b}^T\mathbf{Q}(\boldsymbol{\varrho})^{-1}\mathbf{b}$ is due to the approximation based on the Laplace method.

PQL usually works within the profile likelihood concept. It is distinguished between the estimation of $\boldsymbol{\delta}$ given the plugged-in estimate $\hat{\boldsymbol{\varrho}}$ resulting in the profile-likelihood $l^P(\boldsymbol{\delta}, \hat{\boldsymbol{\varrho}})$ and the estimation of $\boldsymbol{\varrho}$. The PQL method for univariate responses is implemented in the macro `GLIMMIX` and proc `GLMMIX` in SAS (Wolfinger, 1994), in the `glmmPQL` and `gamm` functions of the R-packages `MASS` (Venables and Ripley, 2002) and `mgcv` (Wood, 2006), see also Wolfinger and O'Connell (1993), Littell et al. (1996) and Vonesh (1996).

An implementation of other approaches to fit ordinal random effects models via the Laplace approximation or the adaptive Gauss-Hermite quadrature approximation is available within the R-package `ordinal` (see Christensen, 2010) in the function `clmm`, which uses both approaches to fit cumulative link mixed models. We will focus on the PQL approach and include variable selection by boosting techniques.

3 Boosted Ordinal Random Effects Models

Boosting is a successful and flexible strategy to improve classification procedures. It has been originally developed in the machine learning community. The idea of boosting has become more and more important in the last decade as the issue of estimating high-dimensional models has become more urgent. Since Freund and Schapire (1996) have presented their famous AdaBoost algorithm many other variants in the framework of functional gradient descent optimization have been developed (for example Friedman et al., 2000 or Friedman, 2001). Bühlmann and

Yu (2003) and Bühlmann and Hothorn (2007) extended boosting to generalized linear and additive regression problems based on the L_2-loss, whereas Yuan-Chin et al. (2010) studied on estimating the optimal stopping number of L_2-boosting iterations. Tutz and Binder (2006) proposed a likelihood-based procedure that works for all generalized linear models. The incorporation of random effects has been considered by Tutz and Reithinger (2007) for linear mixed models, first attempts to fit univariate generalized linear models were proposed by Tutz and Groll (2010). Here the case of multivariate mixed models with ordinal response is treated.

3.1 The Basic Algorithm

In the following we present a componentwise boosting algorithm which fits only one linear component x_{itm} of the p-dimensional predictor $\mathbf{x}_{it}^T = (x_{it1}, \ldots, x_{itp})$ at a time. More precisely, a random effects model containing $\gamma_{01}, \ldots, \gamma_{0q}$ and only one linear term $x_{itm}\gamma_m, m = 1, \ldots, p$, is fitted in one iteration step. Let

$$\mathbf{X}_{itm} = \begin{pmatrix} 1 & & & x_{itm} \\ & 1 & & \\ & & \ddots & \vdots \\ & & 1 & x_{itm} \end{pmatrix} = [\mathbf{I}_q, x_{itm}\mathbb{1}_q], \qquad m = 1, \ldots, p,$$

denote the corresponding covariate matrix of observation t in cluster i. Hence we get the predictor

$$\boldsymbol{\eta}_{it.m} = \mathbf{X}_{itm}\boldsymbol{\beta}_m + \mathbf{Z}_{it}\mathbf{b}_i,$$

with $\boldsymbol{\beta}_m^T = (\gamma_{01}, \ldots, \gamma_{0q}, \gamma_m)$ containing only the m-th fixed effect. Furthermore, we will use the notation $\mathbf{X}_{i.m}^T = [\mathbf{X}_{i1m}^T, \ldots, \mathbf{X}_{iT_im}^T]$ for the design matrix of the m-th linear effect in cluster i and analogously we define $\mathbf{X}_{..m}^T = [\mathbf{X}_{1.m}^T, \ldots, \mathbf{X}_{n.m}^T]$ for the whole sample, $m = 1, \ldots, p$. Then for cluster i the predictor that contains only the m-th covariate has the form $\boldsymbol{\eta}_{i..m} = \mathbf{X}_{i.m}\boldsymbol{\beta}_m + \mathbf{Z}_i\mathbf{b}_i$, and for the whole sample we obtain $\boldsymbol{\eta}_{...m} = \mathbf{X}_{..m}\boldsymbol{\beta}_m + \mathbf{Z}\mathbf{b}$.

In the following boosting algorithm the vectors $\boldsymbol{\beta}_m^T = (\gamma_{01}, \ldots, \gamma_{0q}, \gamma_m)$ and $\boldsymbol{\delta}_m^T = (\boldsymbol{\beta}_m^T, \mathbf{b}^T)$ contain only the m-th fixed effect. The algorithm aims at minimizing the likelihood function by iterative fitting of residuals using "weak learners" that fit single candidate predictors. The parameter ν, $0 < \nu \leq 1$, controls the weakness of the learner and is usually set small, say e.g. $\nu = 0.1$. A selection step determines the predictor that is actually updated.

Algorithm `OrdinalBoost`

---

1. *Initialization*
   Compute starting values $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\mathbf{Q}}^{(0)}$ (see Section 3.2.3) and set $\hat{\boldsymbol{\eta}}^{(0)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)} + \mathbf{Z}\hat{\mathbf{b}}^{(0)}$.

2. *Iteration*
   For $l = 1, 2, \ldots, l_{max}$

   (a) *Refitting of residuals*

   i. Computation of parameters
   For $m \in \{1, \ldots, p\}$ the model

   $$g(\boldsymbol{\pi}) = \hat{\boldsymbol{\eta}}^{(l-1)} + \mathbf{X}_{..m}\boldsymbol{\beta}_m + \mathbf{Z}\mathbf{b}$$

   is fitted, where $\hat{\boldsymbol{\eta}}^{(l-1)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(l-1)} + \mathbf{Z}\hat{\mathbf{b}}^{(l-1)}$ is considered a known off-set. Estimation refers to $\boldsymbol{\delta}_m^T = (\boldsymbol{\beta}_m^T, \mathbf{b}^T)$. In order to obtain an additive correction of

the already fitted terms, we use one step in Fisher scoring with starting value $\boldsymbol{\delta}_m = \mathbf{0}$. Therefore Fisher scoring for the m-th component takes the simple form

$$\hat{\boldsymbol{\delta}}_m^{(l)} = (\mathbf{F}_m^{\text{pen}\,(l-1)})^{-1}\mathbf{s}_m^{(l-1)} \tag{5}$$

with penalized pseudo Fisher matrix $\mathbf{F}_m^{\text{pen}\,(l-1)}$ and using the unpenalized version of the penalized score function $\mathbf{s}_m^{\text{pen}\,(l-1)} = \partial l^{\text{pen}}/\partial \boldsymbol{\delta}_m$ (see Section 3.2.1). The variance-covariance components are replaced by their current estimates $\hat{\mathbf{Q}}^{(l-1)}$.

ii. Selection step
Select from $m \in \{1, \ldots, p\}$ the component j that leads to the smallest $AIC_m^{(l)}$ or $BIC_m^{(l)}$ as given in Section 3.2.3. Let the corresponding vector $(\hat{\boldsymbol{\delta}}_j^{(l)})^T$ be denoted by $\left(\hat{\gamma}_{01}^*, \ldots, \hat{\gamma}_{0q}^*, \hat{\gamma}_j^*, (\hat{\mathbf{b}}^*)^T\right)$.

iii. Update step
For $r = 1, \ldots, q$ set

$$\hat{\gamma}_{0r}^{(l)} = \hat{\gamma}_{0r}^{(l-1)} + \hat{\gamma}_{0r}^*,$$

and

$$\hat{\mathbf{b}}^{(l)} = \hat{\mathbf{b}}^{(l-1)} + \hat{\mathbf{b}}^*$$

Set for $m = 1, \ldots, p$

$$\hat{\gamma}_m^{(l)} = \begin{cases} \hat{\gamma}_m^{(l-1)} & \text{if } m \neq j \\ \hat{\gamma}_m^{(l-1)} + \nu\,\hat{\gamma}_m^*, & \text{if } m = j, \quad 0 < \nu \leq 1 \end{cases} \tag{6}$$

which yields

$$\hat{\boldsymbol{\delta}}^{(l)} = \left(\hat{\gamma}_{01}^{(l)}, \ldots, \hat{\gamma}_{0q}^{(l)}, \hat{\gamma}_1^{(l)}, \ldots, \hat{\gamma}_p^{(l)}, (\hat{\mathbf{b}}^{(l)})^T\right)^T.$$

With $\mathbf{A} := [\mathbf{X}, \mathbf{Z}]$ update

$$\hat{\boldsymbol{\eta}}^{(l)} = \mathbf{A}\hat{\boldsymbol{\delta}}^{(l)}$$

(b) *Computation of variance-covariance components*
Estimates of $\hat{\mathbf{Q}}^{(l)}$ are obtained as approximate REML-type estimates or alternative methods (see Section 3.2.2)

3.2 Computational Details of `OrdinalBoost`

In the following we give a more detailed description of the single steps of the `OrdinalBoost` algorithm. First the derivation of the score function and the Fisher matrix are described. Then we present two estimation techniques for the variance-covariance components, give the details of the computation of the starting values and explain the selection procedure.

3.2.1 Score Function and Fisher Matrix

First we specify more precisely the single components which are used in step 2(a) of the algorithm. For $m \in \{1, \ldots, p\}$ the penalized score function $\mathbf{s}_m^{\text{pen}\,(l-1)} = \partial l^{\text{pen}}/\partial \boldsymbol{\delta}_m$, obtained by differentiating the log-likelihood from equation (4), has vector components

$$\begin{aligned} \mathbf{s}_{\boldsymbol{\beta}m}^{\text{pen}\,(l-1)} &= \sum_{i=1}^{n}\sum_{t=1}^{T_i} \mathbf{X}_{itm}^T \mathbf{D}_{it}\boldsymbol{\Sigma}_{it}^{-1}(\mathbf{y}_{it} - \hat{\boldsymbol{\pi}}_{it}), \\ \mathbf{s}_{im}^{\text{pen}\,(l-1)} &= \sum_{t=1}^{T_i} \mathbf{Z}_{it}^T \mathbf{D}_{it}\boldsymbol{\Sigma}_{it}^{-1}(\mathbf{y}_{it} - \hat{\boldsymbol{\pi}}_{it}) - \mathbf{Q}^{-1}\hat{\mathbf{b}}_i^{(l-1)}, \quad i = 1, \ldots, n, \end{aligned}$$

with $\mathbf{D}_{it} = \partial h(\hat{\boldsymbol{\eta}}_{it})/\partial \boldsymbol{\eta}, \boldsymbol{\Sigma}_{it} = cov(\mathbf{y}_{it})$, and $\hat{\boldsymbol{\pi}}_{it} = h(\hat{\boldsymbol{\eta}}_{it})$ evaluated at previous fit $\hat{\boldsymbol{\eta}}_{it} = \mathbf{A}_{it}\hat{\boldsymbol{\delta}}^{(l-1)}$, where $\mathbf{A}_{it} := [\mathbf{X}_{it}, \mathbf{Z}_{it}]$. The vector $\mathbf{s}_{\boldsymbol{\beta}m}^{\text{pen}\,(l-1)}$ has dimension $q+1$, while the vectors $\mathbf{s}_{im}^{\text{pen}\,(l-1)}$ are of dimension s. Note that $\mathbf{s}_{m}^{\text{pen}\,(l-1)}$ could be seen as a penalized score function because of the term $\mathbf{Q}^{-1}\hat{\mathbf{b}}_i^{(l-1)}$.
The penalized pseudo-Fisher matrix $\mathbf{F}_m^{\text{pen}\,(l-1)}$, $m \in \{1, \ldots, p\}$, which is partitioned into

$$\mathbf{F}_m^{\text{pen}\,(l-1)} = \begin{bmatrix} \mathbf{F}_{\boldsymbol{\beta}\boldsymbol{\beta}m} & \mathbf{F}_{\boldsymbol{\beta}1m} & \mathbf{F}_{\boldsymbol{\beta}2m} & \ldots & \mathbf{F}_{\boldsymbol{\beta}nm} \\ \mathbf{F}_{1\boldsymbol{\beta}m} & \mathbf{F}_{11m} & & & 0 \\ \mathbf{F}_{2\boldsymbol{\beta}r} & & \mathbf{F}_{22m} & & \\ \vdots & & & \ddots & \\ \mathbf{F}_{n\boldsymbol{\beta}m} & 0 & & & \mathbf{F}_{nnm} \end{bmatrix},$$

has components

$$\begin{aligned} \mathbf{F}_{\boldsymbol{\beta}\boldsymbol{\beta}m} &= -E\left(\frac{\partial^2 l^{\text{pen}}}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^T}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \mathbf{X}_{itm}^T \mathbf{D}_{it}\boldsymbol{\Sigma}_{it}^{-1}\mathbf{D}_{it}^T\mathbf{X}_{itm}, \\ \mathbf{F}_{\boldsymbol{\beta}im} &= \mathbf{F}_{i\boldsymbol{\beta}m}^T = -E\left(\frac{\partial^2 l^{\text{pen}}}{\partial \boldsymbol{\beta}_m \partial \mathbf{b}_i^T}\right) = \sum_{t=1}^{T_i} \mathbf{X}_{itm}^T \mathbf{D}_{it}\boldsymbol{\Sigma}_{it}^{-1}\mathbf{D}_{it}^T\mathbf{Z}_{it}, \\ \mathbf{F}_{iim} &= -E\left(\frac{\partial^2 l^{\text{pen}}}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}\right) = \sum_{t=1}^{T_i} \mathbf{Z}_{it}^T \mathbf{D}_{it}\boldsymbol{\Sigma}_{it}^{-1}\mathbf{D}_{it}^T\mathbf{Z}_{it} + \mathbf{Q}^{-1}, \end{aligned}$$

with $\mathbf{D}_{it} = \partial h(\hat{\boldsymbol{\eta}}_{it})/\partial \boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_{it} = cov(\mathbf{y}_{it})$ again evaluated at previous fit $\hat{\boldsymbol{\eta}}_{it} = \mathbf{A}_{it}\hat{\boldsymbol{\delta}}^{(l-1)}$.

3.2.2 Variance-Covariance Components

In this section we present two different ways how to update the variance-covariance matrix $\mathbf{Q}$ in step 2(b) of the algorithm. Breslow and Clayton (1993) recommended to estimate the variance by maximizing the profile likelihood that is associated with the normal theory model. By replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ we maximize

$$\begin{aligned} l(\mathbf{Q}_\mathbf{b}) &= -\frac{1}{2}\log(|\mathbf{V}(\hat{\boldsymbol{\delta}})|) - \frac{1}{2}\log(|\mathbf{X}^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}})\mathbf{X}|) \\ &\quad -\frac{1}{2}(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}) - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}})(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}) - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{7}$$

with respect to $\mathbf{Q}_\mathbf{b}$, using the pseudo-observations $\tilde{\boldsymbol{\eta}}(\boldsymbol{\delta}) = \mathbf{A}\boldsymbol{\delta} + \mathbf{D}^{-1}(\boldsymbol{\delta})(\mathbf{y} - \boldsymbol{\pi}(\boldsymbol{\delta}))$ and with matrices $\mathbf{V}(\boldsymbol{\delta}) = \mathbf{W}^{-1}(\boldsymbol{\delta}) + \mathbf{Z}\mathbf{Q}_\mathbf{b}\mathbf{Z}^T$, $\mathbf{W}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta})\mathbf{D}(\boldsymbol{\delta})^T$ and with block-diagonal matrices $\mathbf{Q}_\mathbf{b} = diag(\mathbf{Q}, \ldots, \mathbf{Q})$, $\mathbf{D} = diag(\mathbf{D}_{11}, \ldots, \mathbf{D}_{1T_i}, \mathbf{D}_{21}, \ldots, \mathbf{D}_{2T_2}, \ldots, \mathbf{D}_{nT_n})$ and $\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{1T_i}, \boldsymbol{\Sigma}_{21}, \ldots, \boldsymbol{\Sigma}_{2T_2}, \ldots, \boldsymbol{\Sigma}_{nT_n})$. Having calculated $\hat{\boldsymbol{\delta}}^{(l)}$ in the l-th boosting iteration, we obtain the estimator $\hat{\mathbf{Q}}_\mathbf{b}^{(l)}$, which is an approximate REML-type estimate for $\mathbf{Q}_\mathbf{b}$.

An alternative estimate, which can be derived as an approximate EM algorithm, uses the posterior mode estimates and posterior curvatures. One derives $(\mathbf{F}^{\text{pen}\,(l)})^{-1}$, the inverse of the penalized pseudo Fisher matrix of the full model using the posterior mode estimates $\hat{\boldsymbol{\delta}}^{(l)}$ to obtain the posterior curvatures $\hat{\mathbf{V}}_{ii}^{(l)}$. Now compute $\hat{\mathbf{Q}}^{(l)}$ by

$$\hat{\mathbf{Q}}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{V}}_{ii}^{(l)} + \hat{\mathbf{b}}_i^{(l)}(\hat{\mathbf{b}}_i^{(l)})^T). \tag{8}$$

In general, the $\mathbf{V}_{ii}$ are derived via the formula

$$\mathbf{V}_{ii} = \mathbf{F}_{ii}^{-1} + \mathbf{F}_{ii}^{-1}\mathbf{F}_{i\boldsymbol{\beta}}(\mathbf{F}_{\boldsymbol{\beta}\boldsymbol{\beta}} - \sum_{i=1}^{n}\mathbf{F}_{\boldsymbol{\beta}i}\mathbf{F}_{ii}^{-1}\mathbf{F}_{i\boldsymbol{\beta}})^{-1}\mathbf{F}_{\boldsymbol{\beta}i}\mathbf{F}_{ii}^{-1},$$

where $\mathbf{F}_{\boldsymbol{\beta}\boldsymbol{\beta}}, \mathbf{F}_{i\boldsymbol{\beta}}, \mathbf{F}_{ii}$ are the elements of the penalized pseudo Fisher matrix $\mathbf{F}^{\text{pen}}$ of the full model, for details see for example Tutz and Hennevogl (1996) or Fahrmeir and Tutz (2001). Anderson and Hinde (1988) demonstrated that the principle of the EM approach is generally applicable for generalized linear mixed models.

3.2.3 Starting Values, Hat Matrix and Selection in `OrdinalBoost`

We compute the starting values $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\mathbf{Q}}^{(0)}$ from step 1. of the `OrdinalBoost` algorithm by fitting a simple global intercept model with random effects given by

$$g(\pi_{itr}) = \gamma_{0r} + \mathbf{z}_{it}^T \mathbf{b}_i, \quad i = 1, \ldots, n;\ t = 1, \ldots, T_i;\ r = 1, \ldots, q. \tag{9}$$

This can be done for example by using the R-function `glmmPQL` (Wood, 2006) from the `MASS` library (Venables and Ripley, 2002).

For the derivation of information criteria one has to find the complexity of the fitted model. Following Hastie and Tibshirani (1990) the effective degrees of freedom are determined by the trace of the corresponding hat matrix. Therefore the hat matrix corresponding to the l-th boosting step for the m-th component has to be derived (see also Tutz and Groll, 2010; Tutz and Binder, 2006; Leitenstorfer, 2008).

Let $\mathbf{A}_{..m} := [\mathbf{X}_{..m}, \mathbf{Z}]$ and $\mathbf{K} = diag(0, \ldots, 0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$ be a block-diagonal penalty matrix with a diagonal of $q+1$ zeros corresponding to the q different intercepts $\gamma_{01}, \ldots, \gamma_{0q}$ and the m-th fixed effect γ_m and then n times the matrix $\mathbf{Q}^{-1}$. Then the Fisher matrix $\mathbf{F}_m^{\text{pen}\,(l-1)}$ and the score vector $\mathbf{s}_m^{\text{pen}\,(l-1)}$ are given in closed form as

$$\mathbf{F}_m^{\text{pen}\,(l-1)} = \mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{A}_{..m} + \mathbf{K}$$

and

$$\mathbf{s}_m^{\text{pen}\,(l-1)} = \mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{D}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-1)}) - \mathbf{K}\hat{\boldsymbol{\delta}}_m^{(l-1)},$$

where $\mathbf{W}_l, \mathbf{D}_l, \boldsymbol{\Sigma}_l$ and $\hat{\boldsymbol{\pi}}^{(l-1)}$ are evaluated at the previous fit $\hat{\boldsymbol{\eta}}_{it} = \mathbf{A}\hat{\boldsymbol{\delta}}^{(l-1)}$. For $m = 1, \ldots, p$ the refit in the l-th iteration obtained by a single Fisher scoring step (5) is given by

$$\begin{aligned} \hat{\boldsymbol{\delta}}_m^{(l)} &= (\mathbf{F}_m^{\text{pen}\,(l-1)})^{-1} \mathbf{s}_m^{(l-1)} \\ &= \left(\mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{A}_{..m} + \mathbf{K} \right)^{-1} \mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{D}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-1)}). \end{aligned}$$

We define the predictor corresponding to the m-th refit in the l-th iteration step as

$$\hat{\boldsymbol{\eta}}_{...m}^{(l)} := \hat{\boldsymbol{\eta}}^{(l-1)} + \mathbf{A}_{..m} \boldsymbol{\Psi}_\nu \hat{\boldsymbol{\delta}}_m^{(l)},$$

where $\boldsymbol{\Psi}_\nu = diag(1, \ldots, 1, \nu, 1, \ldots, 1)$ is a diagonal-matrix ensuring that the update of the m-th fixed effect is "weak". Next, we can write

$$\begin{aligned} \hat{\boldsymbol{\eta}}_{...m}^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)} &= \mathbf{A}_{..m} \boldsymbol{\Psi}_\nu \hat{\boldsymbol{\delta}}_m^{(l)} \\ &= \mathbf{A}_{..m} \boldsymbol{\Psi}_\nu \left(\mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{A}_{..m} + \mathbf{K} \right)^{-1} \mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{D}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-1)}). \end{aligned}$$

Taylor approximation of first order $h(\hat{\boldsymbol{\eta}}) \approx h(\boldsymbol{\eta}) + \frac{\partial h(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ yields

$$\begin{aligned} \hat{\boldsymbol{\pi}}_{...m}^{(l)} &\approx \hat{\boldsymbol{\pi}}^{(l-1)} + \mathbf{D}_l (\hat{\boldsymbol{\eta}}_{...m}^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)}), \\ \hat{\boldsymbol{\eta}}_{...m}^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)} &\approx \mathbf{D}_l^{-1} (\hat{\boldsymbol{\pi}}_{...m}^{(l)} - \hat{\boldsymbol{\pi}}^{(l-1)}), \end{aligned}$$

and therefore

$$\mathbf{D}_l^{-1} (\hat{\boldsymbol{\pi}}_{...m}^{(l)} - \hat{\boldsymbol{\pi}}^{(l-1)}) \approx \mathbf{A}_{..m} \boldsymbol{\Psi}_\nu \left(\mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{A}_{..m} + \mathbf{K} \right)^{-1} \mathbf{A}_{..m}^T \mathbf{W}_l \mathbf{D}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-1)}).$$

Multiplication with $\mathbf{W}_l^{1/2}$ and using $\mathbf{W}^{1/2}\mathbf{D}^{-1} = \boldsymbol{\Sigma}^{-1/2}$ yields

$$\boldsymbol{\Sigma}_l^{-1/2}(\hat{\boldsymbol{\pi}}_{\ldots m}^{(l)} - \hat{\boldsymbol{\pi}}^{(l-1)}) \approx \tilde{\mathbf{H}}_m^{(l)} \boldsymbol{\Sigma}_l^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-1)}),$$

where $\tilde{\mathbf{H}}_m^{(l)} := \mathbf{W}_l^{1/2}\mathbf{A}_{..m}\boldsymbol{\Psi}_\nu \left(\mathbf{A}_{..m}^T\mathbf{W}_l\mathbf{A}_{..m} + \mathbf{K}\right)^{-1}\mathbf{A}_{..m}^T\mathbf{W}_l^{1/2}$ denotes the usual generalized ridge regression hat-matrix. Defining $\mathbf{M}_m^{(l)} := \boldsymbol{\Sigma}_l^{1/2}\tilde{\mathbf{H}}_m^{(l)}\boldsymbol{\Sigma}_l^{-1/2}$ yields the approximation

$$\begin{aligned}
\hat{\boldsymbol{\pi}}_{\ldots m}^{(l)} &\approx \hat{\boldsymbol{\pi}}^{(l-1)} + \mathbf{M}_m^{(l)}(\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-1)}) \\
&= \hat{\boldsymbol{\pi}}^{(l-1)} + \mathbf{M}_m^{(l)}[(\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-2)}) - (\hat{\boldsymbol{\pi}}^{(l-1)} - \hat{\boldsymbol{\pi}}^{(l-2)})] \\
&\approx \hat{\boldsymbol{\pi}}^{(l-1)} + \mathbf{M}_m^{(l)}[(\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-2)}) - \mathbf{M}_{j_{l-1}}^{(l-1)}(\mathbf{y} - \hat{\boldsymbol{\pi}}^{(l-2)})],
\end{aligned}$$

whereas $j_{l-1} \in \{1, \ldots, p\}$ denotes the index of the component selected in boosting step $l-1$.

The hat matrix corresponding to the global intercept model from equation (9) is

$$\mathbf{M}^{(0)} = \mathbf{A}_0(\mathbf{A}_0^T\mathbf{W}_1\mathbf{A}_0 + \mathbf{K}_0)\mathbf{A}_0^T\mathbf{W}_1,$$

with $\mathbf{A}_0 := [\mathbf{I}_{nq}, \mathbf{Z}]$, whereas $\mathbf{I}_{nq}^T := [\mathbf{I}_q, \ldots, \mathbf{I}_q]$ consists of n identity matrices of dimension q. We also define the block-diagonal penalty matrix $\mathbf{K}_0 := diag(0, \ldots, 0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$, with a diagonal of q zeros corresponding to the q different intercepts $\gamma_{01}, \ldots, \gamma_{0q}$ and then n times the matrix $\mathbf{Q}^{-1}$. As the approximation $\hat{\boldsymbol{\pi}}^{(0)} \approx \mathbf{M}^{(0)}\mathbf{y}$ holds, one obtains

$$\begin{aligned}
\hat{\boldsymbol{\pi}}_{\ldots m}^{(1)} &\approx \hat{\boldsymbol{\pi}}^{(0)} + \mathbf{M}_m^{(1)}(\mathbf{y} - \hat{\boldsymbol{\pi}}^{(0)}) \\
&\approx \mathbf{M}^{(0)}\mathbf{y} + \mathbf{M}_m^{(1)}(\mathbf{I} - \mathbf{M}^{(0)})\mathbf{y}.
\end{aligned}$$

In the following, to indicate that the hat matrices of the former steps have been fixed, let $j_l \in \{1, \ldots, p\}$ denote the index of the component selected in boosting step l. Then we can abbreviate $\mathbf{M}_{j_l} := \mathbf{M}_{j_l}^{(l)}$ for the matrix corresponding to the component that has been selected in the l-th iteration. Further, in a recursive manner, we get

$$\hat{\boldsymbol{\pi}}_{\ldots m}^{(l)} \approx \mathbf{H}_m^{(l)}\mathbf{y},$$

where

$$\begin{aligned}
\mathbf{H}_m^{(l)} &= \mathbf{I} - (\mathbf{I} - \mathbf{M}_m^{(l)})(\mathbf{I} - \mathbf{M}_{j_{l-1}})(\mathbf{I} - \mathbf{M}_{j_{l-2}}) \cdot \ldots \cdot (\mathbf{I} - \mathbf{M}^{(0)}) \\
&= \mathbf{M}_m^{(l)} \prod_{i=0}^{l-1}(\mathbf{I} - \mathbf{M}_{j_i}) + \sum_{k=0}^{l-1}\mathbf{M}_{j_k}\prod_{i=0}^{k-1}(\mathbf{I} - \mathbf{M}_{j_i}) \\
&= \sum_{k=0}^{l}\mathbf{M}_{j_k}\prod_{i=0}^{k-1}(\mathbf{I} - \mathbf{M}_{j_i}),
\end{aligned}$$

is the hat matrix corresponding to the l-th boosting step considering the m-th component, whereas $\mathbf{M}_{j_l} := \mathbf{M}_m^{(l)}$ is not fixed yet.

For a given hat matrix $\mathbf{H}$, we can determine the complexity of our model by the following information criteria:

$$AIC = -2\,l(\hat{\boldsymbol{\pi}}) + 2\,\text{trace}\,(\mathbf{H}), \tag{10}$$

$$BIC = -2\,l(\hat{\boldsymbol{\pi}}) + 2\,\text{trace}\,(\mathbf{H})\,log(n), \tag{11}$$

where

$$l(\hat{\boldsymbol{\pi}}) = \sum_{i=1}^{n} l_i(\hat{\boldsymbol{\pi}}_i) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\hat{\boldsymbol{\pi}}_i) \tag{12}$$

denotes the log-likelihood and $l_i(\hat{\boldsymbol{\pi}}_i)$ the log-likelihood contribution of $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$. Note that the log-likelihood (4) is given with argument $\boldsymbol{\pi}$ instead of $\boldsymbol{\delta}$, that results from the definition of the natural parameter $\boldsymbol{\theta} = \theta(\boldsymbol{\pi})$ in (2) and by using $\boldsymbol{\pi} = h(\boldsymbol{\eta}) = h(\boldsymbol{\eta}(\boldsymbol{\delta}))$.

For multinomial distributed response the log-likelihood has the form

$$\log f(\mathbf{y}_i|\hat{\boldsymbol{\pi}}_i) = \sum_{t=1}^{T_i} y_{it1} \log \hat{\pi}_{it1} + \ldots + y_{itq} \log \hat{\pi}_{itq} + (1 - y_{it1} - \ldots - y_{itq}) \log (1 - \hat{\pi}_{it1} - \ldots - \hat{\pi}_{itq}).$$

Based on (12), the information criteria (10) and (11) used in the l-th boosting step, considering the m-th component, have the form $AIC_m^{(l)} = -2l(\hat{\boldsymbol{\pi}}_{\ldots m}^{(l)}) + 2\,\text{trace}(\mathbf{H}_m^{(l)})$, $BIC_m^{(l)} = -2l(\hat{\boldsymbol{\pi}}_{\ldots m}^{(l)}) + 2\,\text{trace}\,(\mathbf{H}_m^{(l)})\, log(n)$ with $l(\hat{\boldsymbol{\pi}}_{\ldots m}^{(l)}) = \sum_{i=1}^n \log f(\mathbf{y}_i|\hat{\boldsymbol{\pi}}_{i..m}^{(l)})$.

In the l-th step one selects from $m \in \{1, \ldots, p\}$ the component j_l that minimizes $AIC_m^{(l)}$ or $BIC_m^{(l)}$ and obtains $AIC^{(l)} := AIC_{j_l}^{(l)}$.

3.2.4 Stopping Criterion

As common in boosting we choose a large number l_{max} of maximal boosting steps, e.g. $l_{max} = 1000$, and stop the algorithm at iteration l_{max}. Then selection of the optimal number of boosting steps, l_{opt}, is based on 5-fold cross validation by use of the distance measure

$$D_l := \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left| Y_{it} - \arg \min_r \left\{ \hat{\pi}_{it1}^{(l)} + \cdots + \hat{\pi}_{itr}^{(l)} \geq 0.5 \right\} \right|,$$

which uses the estimate in boosting step l. Afterwards we fit the whole data set again using the `OrdinalBoost` algorithm with l_{opt} boosting iterations to obtain the corresponding parameter estimate $\hat{\boldsymbol{\delta}}^{(l_{opt})}$. Finally we fit a model corresponding to the non-zero parameters of $\hat{\boldsymbol{\delta}}^{(l_{opt})}$ by performing simple Fisher scoring and obtain the final estimates $\hat{\boldsymbol{\delta}}, \hat{\mathbf{Q}}$ and the corresponding fit $\hat{\boldsymbol{\pi}}$.

3.3 Simulation Study

In the following we present two simulation studies to check the performance of the `OrdinalBoost` algorithm, one for the cumulative and one for the sequential model. The algorithm is also compared to alternative approaches. We set $\nu = 1$ in all following simulation studies, since in our experience smaller values hardly improve on the results but more boosting steps are needed which increases the computational cost.

3.3.1 Cumulative Model

The underlying model is a random intercept cumulative logit-model with $k = 6$ response categories and the following design:

$$\begin{aligned}
\eta_{itr} &= \sum_{j=1}^{p} \gamma_{0r} + x_{itj}\gamma_j + b_i, \quad r = 1, \ldots, 5, \quad i = 1, \ldots, 20, \quad t = 1, \ldots, 5, \\
P(Y_{it} = 1) &= F(\eta_{it1}), \\
P(Y_{it} = r) &= F(\eta_{itr}) - F(\eta_{it,r-1}), \quad r = 2, \ldots, 5, \\
P(Y_{it} = 6) &= 1 - F(\eta_{it5}).
\end{aligned}$$

F is chosen as the logistic function, $F(u) = exp(u)/(1 + exp(u))$ yielding the *proportional odds model*. We specify the parameter vector $\boldsymbol{\beta}^T = (\gamma_{01}, \gamma_{02}, \gamma_{03}, \gamma_{04}, \gamma_{05}, \boldsymbol{\gamma}^T) = (-2.5, -1.2, 0, 1.2, 2.5, -15, -20, 35, 0, \ldots, 0)$ with $\gamma_j = 0, j = 4, \ldots, 50$. Therefore three predictors are influential. We choose the different settings $p = 3, 5, 10, 20, 50$. For $j = 1, \ldots, 50$ the vectors $\mathbf{x}_{it}^T =$

$(x_{it1}, \ldots, x_{it50})$ follow a uniform distribution within the interval $[-0.09, 0.09]$ and without correlation. The number of observations is determined by $n = 20, T_i := T = 5, i = 1, \ldots, n$. The random intercepts have been specified by $b_i \sim N(0, \sigma_b^2)$ with three different scenarios, $\sigma_b = 0.4, 0.8, 1.6$.

The performance of estimators is evaluated separately for the structural components and the variance. We compare the results of the `OrdinalBoost` algorithm with the results obtained by the R-function `clmm` which is available in the `ordinal` package (Christensen, 2010). It is able to fit cumulative random effects models using Laplace approximation or adaptive Gauss-Hermite quadrature approximation. Both methods (denoted by $\texttt{clmm}_{\texttt{LP}}$ and $\texttt{clmm}_{\texttt{GH}}$) were used.

By averaging across 100 data sets we consider mean squared errors for $\boldsymbol{\beta}$ and σ_b given by

$$\text{mse}_{\boldsymbol{\beta}} := ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2, \qquad \text{mse}_{\sigma_b} := ||\sigma_b - \hat{\sigma}_b||^2.$$

The results of both quantities for different scenarios of σ_b and for different numbers of noise variables can be found in Table 1. Additional information on the performance of the algorithm was collected in *falseneg*, the mean over all 100 simulations of the number of variables $\gamma_j, j = 1, 2, 3$, that were not selected and in *falsepos*, the mean over all 100 simulations of the number of variables $\gamma_j, j = 4, \ldots, p$, that were selected. As the `clmm` function is not able to perform variable selection it always estimates all parameters $\gamma_j, j = 1, \ldots, p$. For the computation of the random effects variance σ_b^2 we used the two estimation techniques (7) and (8) given in Section 3.2.2. The results can be found in the corresponding `OrdinalBoost` (EM) and (REML) columns of Table 1.

Obviously the boosting estimates clearly outperform the estimates of the `clmm` function, especially in those cases where many noise variables are present. For the case $\sigma_b = 1.6, p = 50$ the `clmm` function did not converge in almost half of the simulations for both Laplace approximation and adaptive Gauss-Hermite quadrature approximation. For the remaining simulations huge values of mean squared errors yielded ($\text{mse}_{\boldsymbol{\beta}} \approx 51500$ and $\text{mse}_{\sigma_b} \approx 22$). As a consequence the boxplots of Figure 3 for $p = 50$ are based on those 49 simulations only, where `clmm` did converge for both techniques. In most of the simulations the REML-type estimates of the parameter vector $\boldsymbol{\beta}$ turned out to perform slightly better than the EM-type estimates and vice versa for the estimates of the standard deviations σ_b of the random effects. The corresponding boxplots of $\text{mse}_{\boldsymbol{\beta}}$ are shown in Figures 1-3, for different numbers of noise variables and for different scenarios of σ_b. Figure 4 exemplarily shows the boxplots of the ratios $\log(\text{mse}_{\sigma_b}(\ldots)/\text{mse}_{\sigma_b}(\texttt{clmm}(\text{LP})))$ corresponding to the $\sigma_b = 0.4$ case.

3.3.2 Sequential Model

The underlying model is a random intercept sequential logit-model with $k = 6$ response categories and the following design:

$$\begin{aligned}
\eta_{itr} &= \sum_{j=1}^{p} \gamma_{0r} + x_{itj}\gamma_j + b_i, \quad r = 1, \ldots, 5, \quad i = 1, \ldots, 20, \quad t = 1, \ldots, 5, \\
P(Y_{it} = 1) &= F(\eta_{it1}), \\
P(Y_{it} = r) &= F(\eta_{itr}) \prod_{j=1}^{r-1} \left(1 - F(\eta_{itj})\right), \quad r = 2, \ldots, 5, \\
P(Y_{it} = 6) &= \prod_{j=1}^{5} \left(1 - F(\eta_{itj})\right).
\end{aligned}$$

Again we choose F to be the logistic function $F(u) = exp(u)/(1 + exp(u))$, we choose the same parameter vector $\boldsymbol{\beta}^T = (\gamma_{01}, \gamma_{02}, \gamma_{03}, \gamma_{04}, \gamma_{05}, \boldsymbol{\gamma}^T) = (-2.5, -1.2, 0, 1.2, 2.5, -20, 35, 0, \ldots, 0)$ with $\gamma_j = 0, j = 4, \ldots, 50$ as for the cumulative model and investigate the different settings

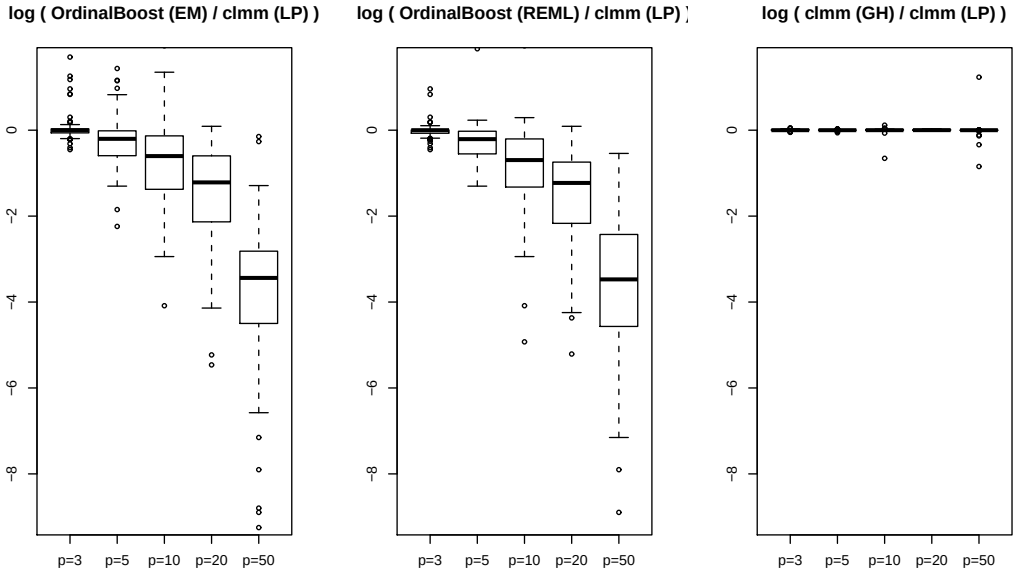

Figure 1: Cumulative logit model: Boxplots of the ratios $\log(\text{mse}_{\boldsymbol{\beta}}(\ldots)/\text{mse}_{\boldsymbol{\beta}}(\texttt{clmm}(\text{LP})))$ for the three different methods, for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 0.4$

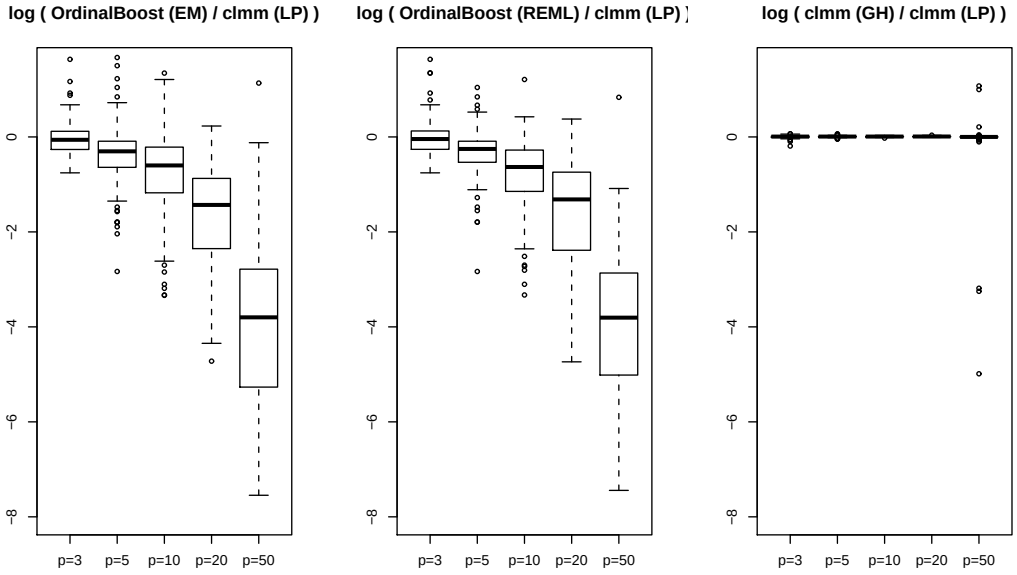

Figure 2: Cumulative logit model: Boxplots of the ratios $\log(\text{mse}_{\boldsymbol{\beta}}(\ldots)/\text{mse}_{\boldsymbol{\beta}}(\texttt{clmm}(\text{LP})))$ for the three different methods, for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 0.8$

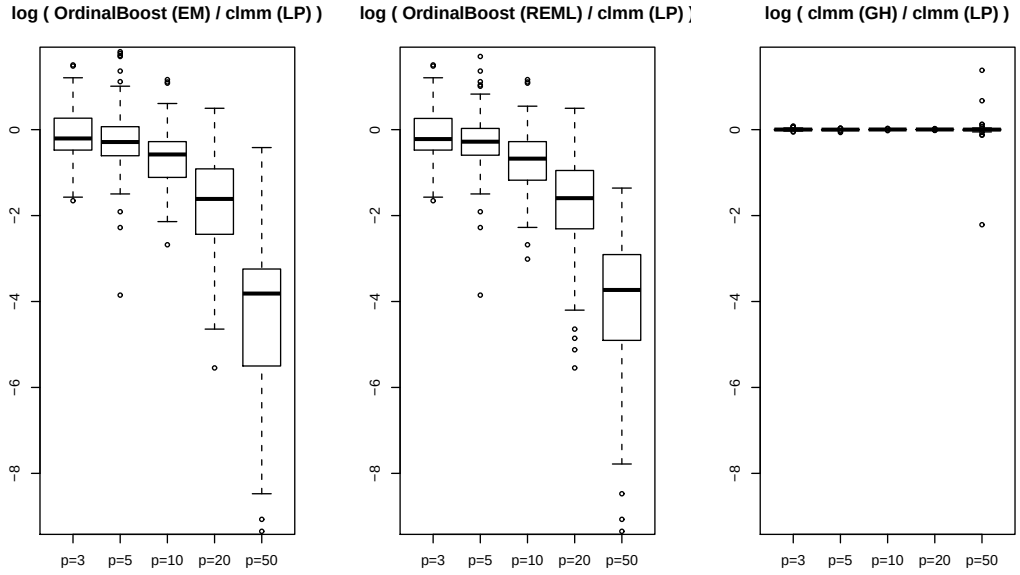

Figure 3: Cumulative logit model: Boxplots of the ratios $\log(\text{mse}_{\boldsymbol{\beta}}(\ldots)/\text{mse}_{\boldsymbol{\beta}}(\texttt{clmm}(\text{LP})))$ for the three different methods, for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 1.6$

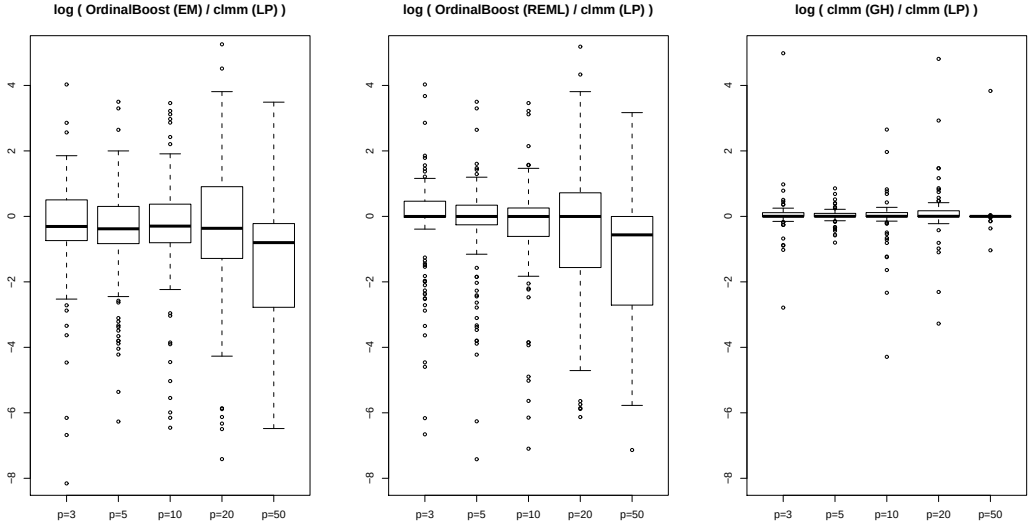

Figure 4: Boxplots of the ratios $\log(\text{mse}_{\sigma_b}(\ldots)/\text{mse}_{\sigma_b}(\texttt{clmm}(\text{LP})))$ for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 0.4$ for the cumulative logit model

$p = 3, 5, 10, 20, 50$. For $j = 1, \ldots, 50$ the vectors $\mathbf{x}_{it}^T = (x_{it1}, \ldots, x_{it50})$ follow the same uniform distribution on the interval $[-0.09, 0.09]$. The number of observations remains $n = 20, T_i := T = 5, i = 1, \ldots, n$ and the random intercepts have been specified by $b_i \sim N(0, \sigma_b^2)$ with the same three different cases $\sigma_b = 0.4, 0.8, 1.6$.

We use the same goodness-of-fit criteria as for the cumulative model to check the performance of our `OrdinalBoost` algorithm and compare with the results obtained from the R-functions `glmmPQL` and `glmer`. Both functions assume a binary response but it is well known that the sequential model can be fitted as a binary response model by defining binary indicators for the transitions (see for example Fahrmeir and Tutz, 2001). The `glmmPQL` function is available in the `MASS` package (Venables and Ripley, 2002). It operates by iteratively calling the R-function `lme` from the `nlme` library and returns the fitted `lme` model object for the working model at convergence. For more details about the `lme` function, see Pinheiro and Bates (2000) and Wood (2006). The `glmer` function available in the `lme4` package (Bates and Maechler, 2010). It features two different methods of approximating the integrals in the log-likelihood function, Laplace and Gauss-Hermite, where we focused on the Gauss-Hermite method in our simulations using 15 quadrature points. In some cases the `glmer` function did not estimate random effects and set $\sigma_b = 0$. As a consequence we derived the mean squared errors in Table 2 only for those cases where `glmer` did estimate random effects. Besides the number of simulations, where no random (n. r.) effects were estimated can be found in the corresponding column. Another function that is able to fit the underlying model is the `glmmML` function supplied with the `glmmML` package (Broström, 2009). The function also features two different methods of approximating the integrals in the log-likelihood function, Laplace and Gauss-Hermite. For the first method the results coincide with the results of the `glmmPQL` routine, so we focused on the Gauss-Hermite method in our simulations. But as the estimates of especially the random effects standard deviation σ_b have been a good deal worse as for the `glmer` function we abstained on presenting its results. All other results are summarized in Table 2.

It is obvious that both parameter and variance estimates of the two boosting methods remain relatively stable when the number of noise variables is increasing, whereas the mean squared errors of the estimates obtained by the two R-functions are drastically deteriorating. This effect becomes distinct especially for the variance estimates. For $p = 50$ the mean squared errors for the `glmmPQL` function explode yielding values of $\text{mse}_{\boldsymbol{\beta}} > 9 \cdot 10^{22}$ and $\text{mse}_{\sigma_b} > 3 \cdot 10^7$. The same happens for the `glmer` function just with a more moderate order of magnitude. The results of $\text{mse}_{\boldsymbol{\beta}}$ are illustrated in the Figures 5 - 7 which show boxplots of the ratios $\log(\text{mse}_{\boldsymbol{\beta}}(\ldots)/\text{mse}_{\boldsymbol{\beta}}(\texttt{glmer}))$ for the different methods, for different numbers of noise variables and for different scenarios of σ_b. Figure 8 exemplarily shows the boxplots of the ratios $\log(\text{mse}_{\sigma_b}(\ldots)/\text{mse}_{\sigma_b}(\texttt{glmer}))$ corresponding to the $\sigma_b = 1.6$ case.

Comparison of boosting procedures yields that the REML estimates of the parameter vector are slightly more stable than the EM estimates, except for a high number of noise variables, e.g. the $p = 50$ case. In this case the `glmmPQL` function did not provide satisfying estimates, which let to the high values of the mean squared errors. In estimating the random effects variance σ_b^2 the REML-type boosting clearly outperforms the EM-type for all numbers of noise variables and in all three scenarios for σ_b, except for the scenario with $p = 50, \sigma_b = 1.6$.

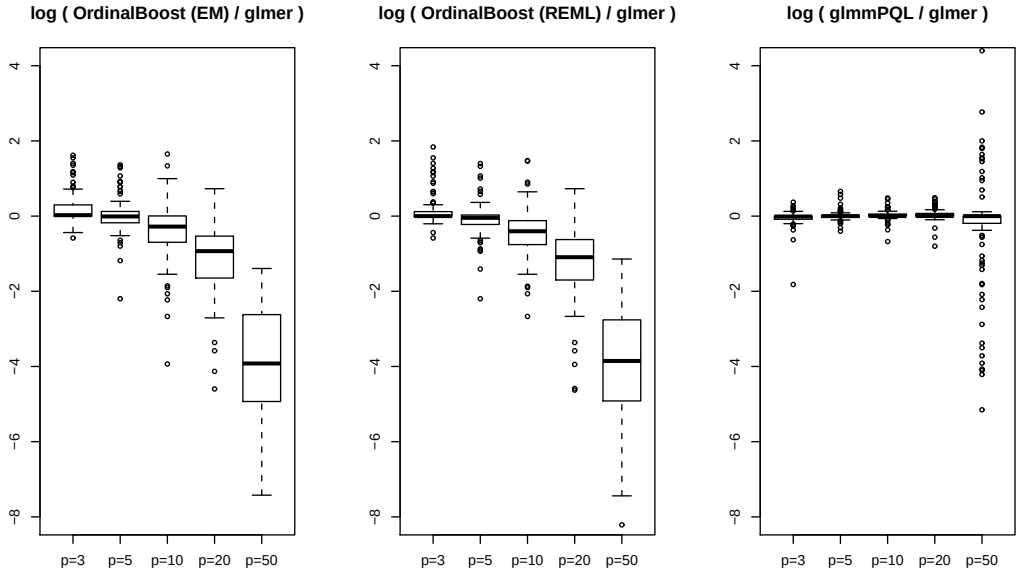

Figure 5: Sequential logit model: Boxplots of the ratios $\log(\text{mse}_{\boldsymbol{\beta}}(\ldots)/\text{mse}_{\boldsymbol{\beta}}(\texttt{glmer}))$ for the three different methods, for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 0.4$

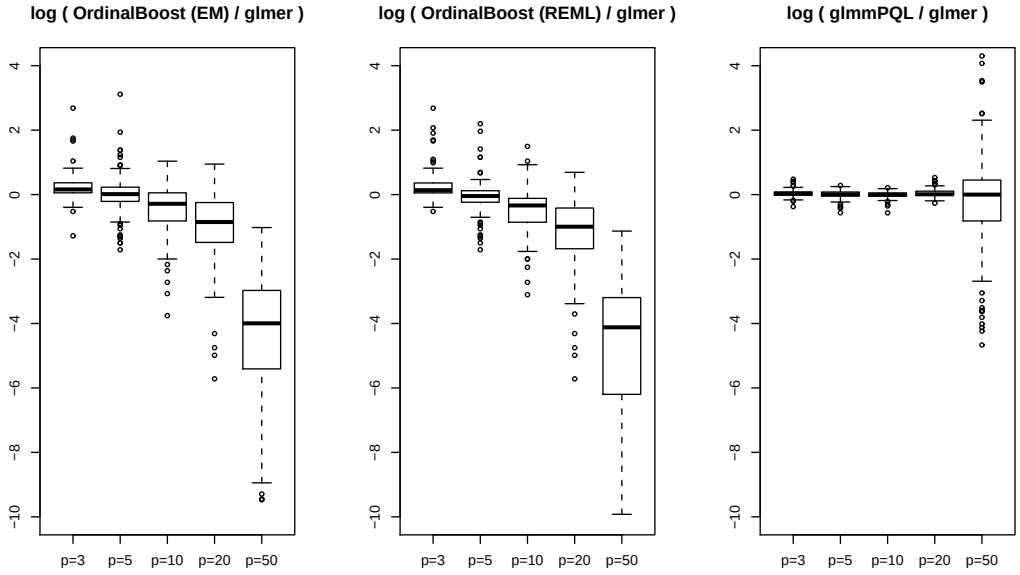

Figure 6: Sequential logit model: Boxplots of the ratios $\log(\text{mse}_{\boldsymbol{\beta}}(\ldots)/\text{mse}_{\boldsymbol{\beta}}(\texttt{glmer}))$ for the three different methods, for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 0.8$

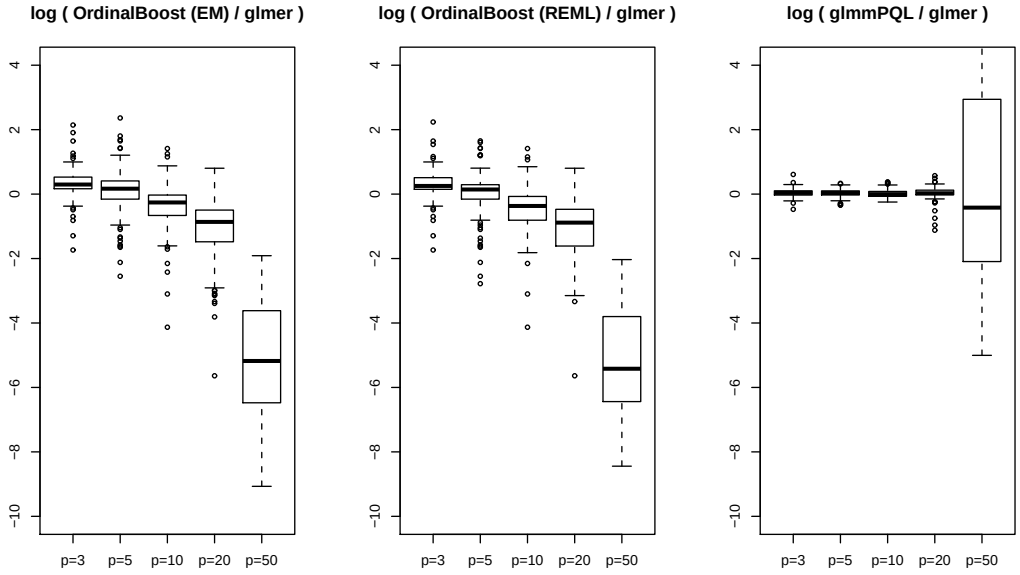

Figure 7: Sequential logit model: Boxplots of the ratios $\log(\text{mse}_{\boldsymbol{\beta}}(\ldots)/\text{mse}_{\boldsymbol{\beta}}(\texttt{glmer}))$ for the three different methods, for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 1.6$

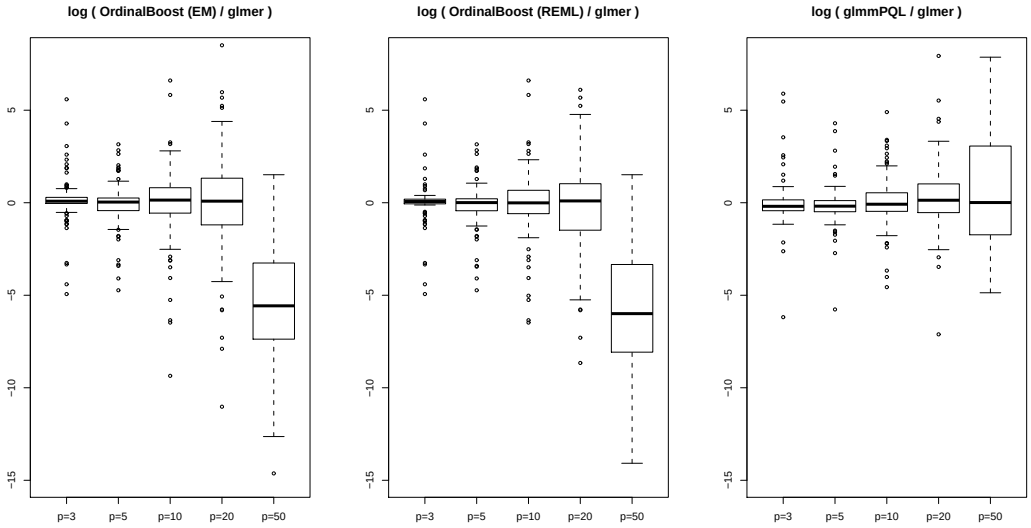

Figure 8: Boxplots of the ratios $\log(\text{mse}_{\sigma_b}(\ldots)/\text{mse}_{\sigma_b}(\texttt{glmer}))$ for $p = 3, 5, 10, 20, 50$ covariables and $\sigma_b = 1.6$ for the sequential logit model

		clmm$_{LP}$		clmm$_{GH}$		OrdinalBoost (EM)				OrdinalBoost (REML)			
σ	p	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	falsepos	falseneg	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	falsepos	falseneg
0.4	3	71.469	0.117	72.056	0.121	80.812	0.101	0	0.08	72.784	0.119	0	0.04
0.4	5	105.876	0.113	106.182	0.116	102.611	0.087	0.23	0.12	86.041	0.110	0.34	0.02
0.4	10	259.863	0.151	259.682	0.156	166.128	0.122	0.96	0.11	152.049	0.138	1.00	0.03
0.4	20	752.915	0.140	758.684	0.148	274.146	0.110	1.68	0.16	248.585	0.119	1.46	0.10
0.4	50	17048.85	5.007	15154.94	3.840	481.533	0.270	2.09	0.33	607.036	0.331	2.86	0.24
0.8	3	80.197	0.128	82.707	0.130	82.707	0.107	0	0.06	83.098	0.116	0	0.07
0.8	5	140.788	0.176	141.894	0.177	118.900	0.134	0.38	0.10	115.574	0.142	0.51	0.06
0.8	10	264.126	0.228	266.355	0.230	156.264	0.131	0.97	0.12	160.873	0.148	1.02	0.09
0.8	20	769.825	0.278	777.261	0.285	212.858	0.131	1.36	0.12	233.877	0.149	1.44	0.14
0.8	50	22212.51	10.654	24125.34	9.183	519.264	0.312	2.02	0.21	443.368	0.239	1.75	0.22
1.6	3	89.293	0.175	89.460	0.173	90.226	0.142	0	0.08	87.397	0.143	0	0.07
1.6	5	142.667	0.202	142.887	0.203	129.464	0.158	0.37	0.15	121.091	0.156	0.40	0.11
1.6	10	375.364	0.404	377.015	0.407	211.074	0.278	1.15	0.14	198.827	0.273	1.05	0.11
1.6	20	954.661	0.531	958.653	0.544	250.771	0.225	1.46	0.24	256.109	0.233	1.37	0.22
1.6	50	-	-	-	-	447.677	0.296	1.44	0.39	651.837	0.395	2.09	0.39

Table 1: Cumulative mixed logit model with `clmm` and boosting (`OrdinalBoost`).

		glmmPQL		glmer			OrdinalBoost (EM)				OrdinalBoost (REML)			
σ	p	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	n. r.	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	falsepos	falseneg	mse$_{\boldsymbol{\beta}}$	mse$_{\sigma_b}$	falsepos	falseneg
0.4	3	93.969	0.090	97.257	0.080	23	128.500	0.120	0	0.09	124.608	0.086	0	0.12
0.4	5	130.650	0.096	126.963	0.074	20	149.034	0.117	0.18	0.12	133.235	0.078	0.11	0.09
0.4	10	210.170	0.149	198.403	0.094	18	164.660	0.129	0.53	0.09	159.413	0.084	0.37	0.11
0.4	20	528.961	0.161	500.154	0.098	27	211.668	0.109	0.98	0.18	189.044	0.073	0.83	0.14
0.4	50	-	-	121726.2	48.197	42	350.264	0.197	1.44	0.33	313.535	0.138	1.37	0.32
0.8	3	123.089	0.132	119.232	0.111	6	156.287	0.197	0	0.13	166.620	0.117	0	0.20
0.8	5	130.051	0.118	129.748	0.115	3	148.592	0.231	0.22	0.09	137.151	0.106	0.23	0.10
0.8	10	184.792	0.101	187.460	0.103	2	157.899	0.231	0.43	0.17	154.056	0.091	0.39	0.18
0.8	20	548.157	0.223	521.716	0.146	9	240.418	0.244	1.08	0.24	214.812	0.131	0.79	0.25
0.8	50	-	-	119067.1	67.907	26	351.237	0.353	1.26	0.30	240.537	0.132	0.80	0.27
1.6	3	113.927	0.199	108.230	0.225	0	156.064	0.357	0	0.10	146.997	0.221	0	0.09
1.6	5	152.470	0.229	145.577	0.245	0	207.676	0.289	0.21	0.25	190.985	0.227	0.19	0.19
1.6	10	265.0.78	0.191	255.860	0.182	0	212.369	0.263	0.84	0.17	197.680	0.165	0.74	0.16
1.6	20	764.310	0.575	715.078	0.343	1	292.503	0.304	1.23	0.27	265.698	0.223	1.01	0.32
1.6	50	-	-	121644.5	96.471	9	426.098	0.305	1.81	0.40	375.712	0.206	1.07	0.54

Table 2: Sequential mixed logit model with `glmmPQL`, `glmer` and boosting (`OrdinalBoost`)

4 Applications to Real Data

In the following sections we apply our boosting method on different real data sets and compare the results with other approaches. It should be noted that fixed effects of categorical predictors and predictors that include polynomial terms are updated blockwise in step (6) of the boosting algorithm. Moreover, we derive the optimal number of boosting steps by selecting the step l_{opt} for which $AIC^{(l)}$ is lowest instead of using cross-validation.

4.1 Recovery Data

The data set was published by Davis (1991). In the study the impact of different doses of an anesthetic is analyzed for 60 children. As soon as the children enter the anesthetic recovery room after a surgery their level of "awakeness" is measured followed by three further measurements after 5, 15 and 30 minutes. The level of "awakeness" is given on a spectrum ranging from 0 (sleeping) to 6 (awake). For each child the categorical influence variable "Dose" (dosage of the anesthetic; 15, 20, 25 or 30 mg/kg) as well as the metric influence variables "Age" (in month) and "Duration of the surgery" (in minutes) have been observed. Finally we include another categorical variable, the "Number of Replication". We use a cumulative random intercept model to fit the data allowing "Age" to have a non-linear effect by including the variable "Age^2". The corresponding linear predictor is

$$\begin{aligned}\eta_{itr} &= \gamma_{0r} + \text{Dose1}_{it}\gamma_1 + \text{Dose2}_{it}\gamma_2 + \text{Dose3}_{it}\gamma_3 + \text{Duration}_{it}\gamma_4 + \text{Age}_{it}\gamma_5 + \text{Age}^2_{it}\gamma_6 \\ &\quad + \text{Replication1}_{it}\gamma_7 + \text{Replication2}_{it}\gamma_8 + \text{Replication2}_{it}\gamma_9 + b_i, \quad r = 1, \ldots, 6.\end{aligned}$$

"Dose4" (30 mg/kg) and "Replication4" are used as reference categories. Figure 9 shows the corresponding coefficient paths for the `OrdinalBoost` algorithm (EM). The covariates have been standardized by dividing them by their empirical standard deviation.

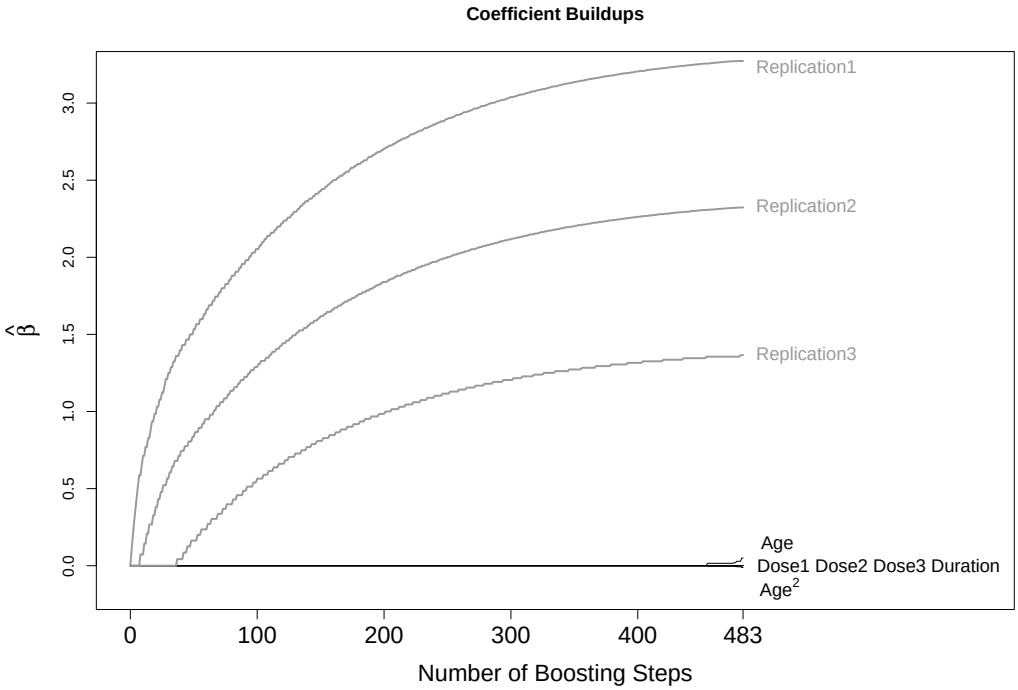

Figure 9: Coefficient buildups of the EM-estimates for standardized covariates, recovery data

Table 3 shows the results of the estimates obtained by the R-function `clmm`, using Laplace (LP) and Gauss-Hermite (GH), as well as the results obtained by our `OrdinalBoost` algorithm. The

estimates of the EM and the REML technique are very similar, with the main difference, that the REML approach didn't select the variables "Age" and "Age^2". For the R-function `clmm` some convergence problems occur and for both methods (LP and GH) the warning message is reported, that the variance-covariance matrix of the parameters is not defined, with the consequence that the standard deviations of the estimates could not be derived. As is seen in Figure 9 the `OrdinalBoost` algorithm using the EM technique does not select the variables "Duration" and "Dose". So in the final Fisher scoring (see Section 3.2.4) a model without these two variables is fitted and both variables are not incorporated into the model.

	`clmmLP`	`clmmGH`	`OrdinalBoost`	
			(EM)	(REML)
Intercept 1	-10.382 (NA)	-10.389 (NA)	-9.252	-8.749
Intercept 2	-6.412 (NA)	-6.407 (NA)	-5.722	-5.249
Intercept 3	-5.333 (NA)	-5.325 (NA)	4.705	-4.244
Intercept 4	-3.767 (NA)	-3.758 (NA)	-3.258	-2.814
Intercept 5	-2.636 (NA)	-2.625 (NA)	-2.198	-1.769
Intercept 6	-1.263 (NA)	-1.247 (NA)	-0.901	-0.493
Dose 1	-1.700 (NA)	-1.725 (NA)	0	0
Dose 2	-1.530 (NA)	-1.559 (NA)	0	0
Dose 3	-0.632 (NA)	-0.642 (NA)	0	0
Duration	0.016 (NA)	0.016 (NA)	0	0
Age	0.033 (NA)	0.033 (NA)	0.035	0
Age^2	0.001 (NA)	0.001 (NA)	0.002	0
Replication1	6.482 (NA)	6.489 (NA)	5.882	5.805
Replication2	4.719 (NA)	4.727 (NA)	4.301	4.242
Replication3	2.992 (NA)	2.990 (NA)	2.710	2.677
$\hat{\sigma}_b$	3.444 (NA)	3.511 (NA)	3.259	3.237

Table 3: Estimates for the recovery data

4.2 Forest health Data

The forest health data has been considered in previous studies, for example in Kneib and Fahrmeir (2010) and Kneib et al. (2009). In this application, the health status of beeches at 83 observation plots located in a northern Bavarian forest district has been assessed in visual forest health inventories carried out between 1983 and 2004. Originally, the health status is classified on an ordinal scale, where the nine possible categories denote different degrees of defoliation. Figure 10 shows a histogram of the nine defoliation classes indicating that no trees were observed in the last two categories. Therefore we use the categorical response variable "defoliation" with seven categories by aggregating over the last three categories. In Kneib et al. (2009) a brief description of the covariates in the data set is presented, which can be found in Table 4.

We use a sequential random intercept model to fit the data, allowing "age" and "time" to have not only strictly linear effects by including the variables "age^2" and "time^2" into our model. Kneib et al. (2009) identified also nonlinear effects for canopy density and soil depth, so we also incorporate the variables "canopy^2" and "soil depth^2" for our model. Figure 11 shows the corresponding coefficient paths of the `OrdinalBoost` algorithm with REML approach for standardized covariates. Due to our choice of $\nu = 0.1$ the AIC did still slightly improve at a high number of boosting steps, which also reflects the resistance against overfitting, which is an important trail of boosting procedures. Step 135 was the last step where a new variable ("Elevation") entered the model and from step 267 on only the variables "age" and "age^2" were slightly updated, so the algorithm was stopped at step $l_{max} = 1000$.

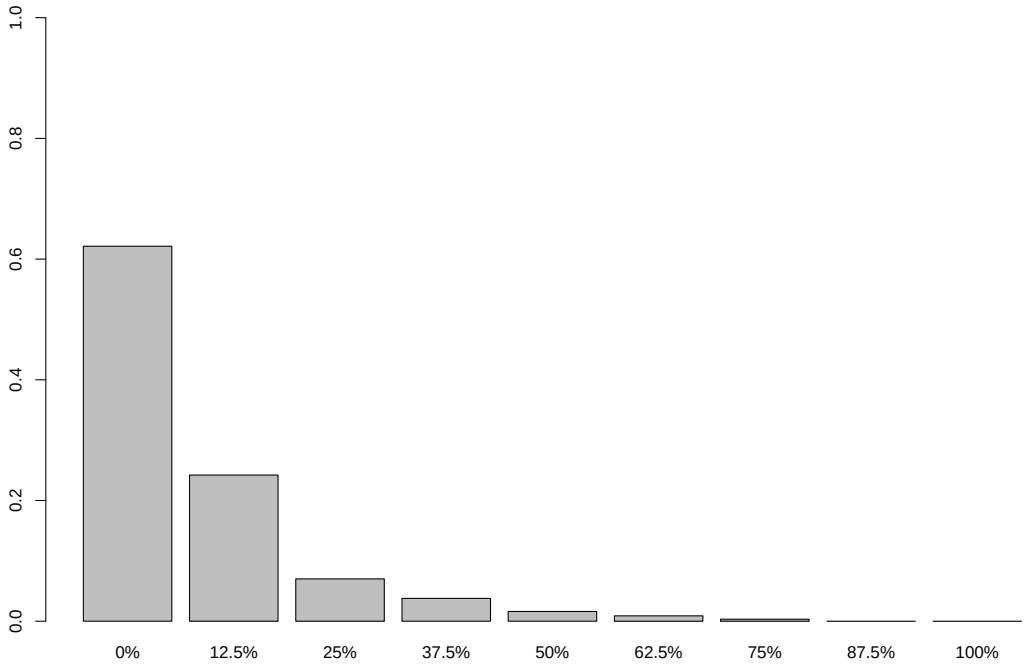

Figure 10: Relative frequencies of the nine defoliation classes for all observation plots and all time points for the forest health data

Covariate	Description
age	age of the tree in years (continuous, $7 \leq$ age ≤ 234)
time	calendar time (continuous, $1983 \leq$ time ≤ 2004)
elevation	elevation above sea level in meters (continuous, $250 \leq$ elevation ≤ 480)
inclination	inclination of slope in percent (continuous, $0 \leq$ inclination ≤ 46)
soil	depth of soil layer in centimeters (continuous, $9 \leq$ soil ≤ 51)
canopy	density of forest canopy in percent (continuous, $0 \leq$ canopy ≤ 1)
stand	type of stand (categorical, 1 = deciduous forest, -1 = mixed forest)
fertilisation	fertilisation (categorical, 1 = yes, -1 = no)
humus	thickness of humus layer in 5 categories (ordinal, higher categories represent higher proportions)
moisture	level of soil moisture (categorical, 1 = moderately dry, 2 = moderately moist, 3 = moist or temporary wet)
saturation	base saturation (ordinal, higher categories indicate higher base saturation)

Table 4: Description of covariates for the forest health data

Table 5 shows the results of the estimates obtained by the R-functions `glmmPQL` and `glmer` as well as the results obtained by our `OrdinalBoost` algorithm. The estimates of the EM and the REML technique are very similar, with the main difference, that the REML approach did not select the variables "time" and "time^2".

Although fitting procedures for the full model still work, the main advantage of the boosting procedure, selection of parameters, becomes obvious. REML boosting deletes 5 variables of the 11 available variables.

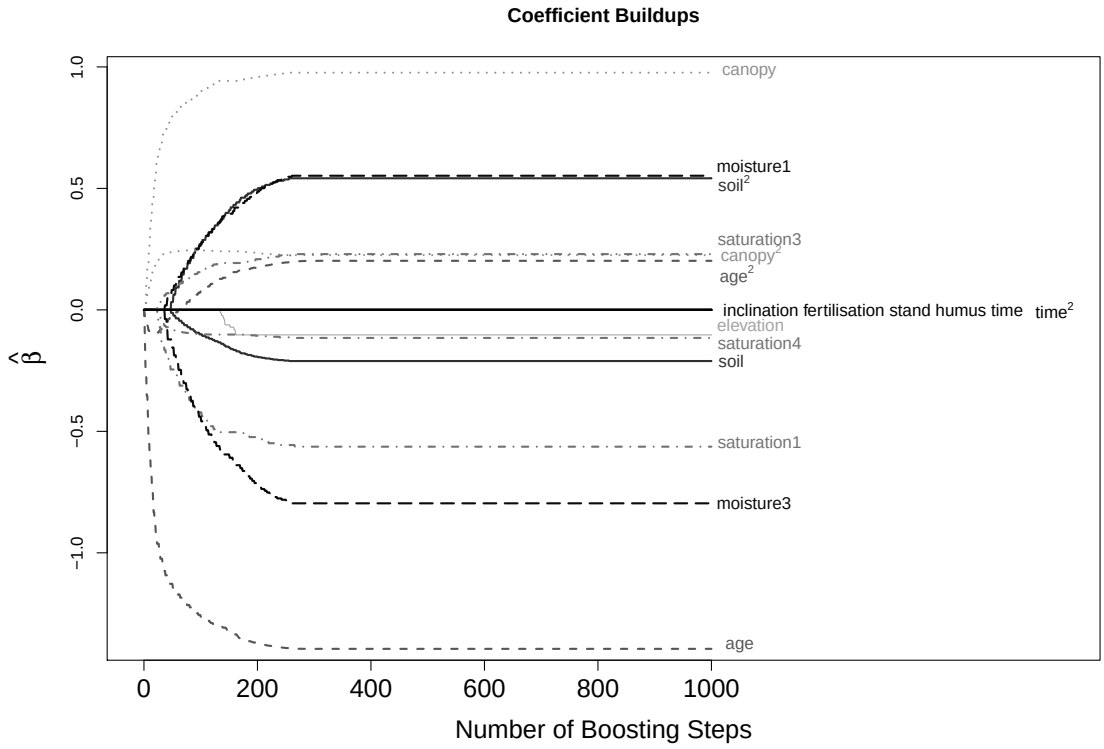

Figure 11: Coefficient buildups of the REML-estimates for standardized covariates, forest health data

5 Concluding Remarks

Procedures for the fitting of binary and ordinal mixed models in high-dimensional designs were proposed and examined. The selection procedures work quite stable and allow to select the influential variables from a set of variables which includes irrelevant ones. They also work in cases where common methods that are unable to select predictors fail.

The used method is an adaptation of likelihood-based boosting to generalized linear mixed models. Alternatively one could use L_1-penalty techniques by maximizing the penalized marginal likelihood. A procedure of that type has been proposed more recently for the semi-parametric linear mixed model (see Ni et al., 2010). The generalization to generalized linear models seems not yet available.

One should also mention an alternative boosting scheme that is available in the `mboost` package (see Hothorn et al., 2010 and Bühlmann and Hothorn, 2007). The package provides a variety of gradient boosting families to specify loss functions and the corresponding risk functions to be optimized. It has recently been extended to families with an additional scale parameter, for example the `PropOdds()` family leads to the (fixed effects) proportional odds model, see Schmid and Hothorn (2008) and Schmid et al. (2010). The `gamboost` function from that package also allows to model heterogeneity in repeated measurements, but fits a fixed parameter model. No distribution assumption for the random effects is used and thus no estimates for the variance of the random effects can be derived (see also Kneib et al., 2009). Therefore modeling and estimates are not comparable to the generalized mixed model approach that has been proposed here.

	glmmPQL	glmer	OrdinalBoost (EM)	OrdinalBoost (REML)
Intercept 1	2.892 (2.552)	3.390 (2.717)	0.113	-0.028
Intercept 2	5.017 (2.554)	5.571 (2.720)	1.882	1.737
Intercept 3	5.726 (2.559)	6.297 (2.725)	2.372	2.232
Intercept 4	7.061 (2.567)	7.666 (2.733)	3.567	3.427
Intercept 5	7.824 (2.581)	8.447 (2.748)	4.175	4.033
Intercept 6	9.263 (2.623)	9.916 (2.791)	5.661	5.541
time	0.024 (0.011)	0.025 (0.011)	0.015	0
time^2	0.002 (0.001)	0.002 (0.002)	-0.001	0
inclination	0.013 (0.030)	0.013 (0.032)	0	0
elevation	-0.006 (0.006)	-0.007 (0.006)	-0.002	-0.001
soil	0.035 (0.038)	0.040 (0.041)	-0.021	-0.019
soil^2	-0.003 (0.002)	-0.003 (0.003)	0.002	0.002
fertilisation	2.868 (1.017)	3.111 (1.095)	0	0
age	-0.055 (0.006)	-0.058 (0.006)	-0.024	-0.024
age^2	-0.000 (0.000)	-0.001 (0.000)	-0.000	-0.000
canopy	2.692 (0.496)	2.759 (0.519)	4.092	3.861
canopy^2	0.136 (1.454)	0.387 (1.529)	3.23	2.447
stand	-0.487 (0.482)	-0.576 (0.512)	0	0
saturation1	-0.632 (0.631)	-0.680 (0.671)	-0.775	-0.768
saturation3	-0.124 (0.677)	-0.168 (0.722)	0.413	0.439
saturation4	-0.060 (0.836)	-0.095 (0.892)	-0.128	-0.162
humus0	0.397 (0.125)	0.409 (0.131)	0	0
humus2	-0.069 (0.102)	-0.070 (0.106)	0	0
humus3	-0.277 (0.121)	-0.284 (0.127)	0	0
humus4	-0.239 (0.168)	-0.248 (0.176)	0	0
moisture1	0.623 (0.650)	0.763 (0.698)	0.748	0.747
moisture3	-1.319 (0.459)	-1.468 (0.490)	-0.747	-0.723
$\hat{\sigma}_b$	2.214 (0.961)	2.358	1.448	1.421

Table 5: Estimates for the forest health data

References

Anderson, D. A. and J. P. Hinde (1988). Random effects in generalized linear models and the EM algorithm. *Communications in Statistics-Theory and Methods 17*(11).

Bates, D. and M. Maechler (2010). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-34.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

Breslow, N. E. and X. Lin (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91.

Broström, G. (2009). *glmmML: Generalized linear models with clustering*. R package version 0.81-6.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*, 477–522.

Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Christensen, R. H. B. (2010). ordinal—regression models for ordinal data. R package version 2010.05-17 http://www.cran.r-project.org/package=ordinal/.

Davis, C. S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine 10*, 1959–1980.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer-Verlag.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalize likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Freund, Y. and R. E. Schapire (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148–156. San Francisco, CA: Morgan Kaufmann.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics 29*, 337–407.

Friedman, J. H., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics 28*, 337–407.

Hartzel, J., A. Agresti, and B. Caffo (2001). Multinomial logit random effects models. *Statistical Modelling 1*, 81–102.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner (2010). mboost: Model-based boosting. R package version 2.0-6 http://CRAN.R-project.org/package=mboost/.

Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics 39*, 74–85.

Kneib, T. and L. Fahrmeir (2010). A Space-Time Study on Forest Health. In R. E. Chandler and M. Scott (Eds.), *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. Wiley (to appear).

Kneib, T., T. Hothorn, and G. Tutz (2009). Variable selection and model choice in geoadditive regression. *Biometrics 65*, 626–634.

Leitenstorfer, F. (2008). *Boosting in Nonparametric Regression: Constrained and Unconstrained Modeling Approaches*. München: Verlag Dr. Hut.

Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*, 1007–1016.

Littell, R., G. Milliken, W. Stroup, and R. Wolfinger (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institue Inc.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*, 109–142.

Ni, X., D. Zhang, and H. H. Zhang (2010). Variable Selection for Semiparametric Mixed Models in Longitudinal Studies. *Biometrics 66*, 79–88.

Park, M. Y. and T. Hastie (2006). An l1 regularization-path algorithm for generalized linear models. *Preprint, Department of Statistics, Stanford University*.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Schmid, M. and T. Hothorn (2008). Flexible boosting of accelerated failure time models. *BMC Bioinformatics*, 9–296.

Schmid, M., S. Potapov, A. Pfahlberg, and T. Hothorn (2010). Estimation and regularization techniques for regression models with multidimensional prediction functions. *Statistics and Computing 20*.

Tutz, G. and H. Binder (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics 62*, 961–971.

Tutz, G. and A. Groll (2010). Generalized Linear Mixed Models Based on Boosting. In T. Kneib and G. Tutz (Eds.), *Statistical Modelling and Regression Structures - Festschrift in the Honour of Ludwig Fahrmeir*. Physica.

Tutz, G. and W. Hennevogl (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis 22*, 537–557.

Tutz, G. and F. Reithinger (2007). Flexible semiparametric mixed models. *Statistics in medicine 26*, 2872–2900.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Vonesh, E. F. (1996). A note on the use of laplace's approximatio for nonlinear mixed-effects models. *Biometrika 83*, 447–452.

Wolfinger, R. and M. O'Connell (1993). Generalized linear mixed models; a pseudolikelihood approach. *Journal Statist. Comput. Simulation 48*, 233–243.

Wolfinger, R. W. (1994). Laplace's approximation for nonlinear mixed models. *Biometrika 80*, 791–795.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Yuan-Chin, I. C., H. Yufen, and H. Yu-Pai (2010). Early stopping in L2 boosting. *Computational Statistics & Data Analysis 54*, 2203–2213.